NASA-CR-201009

A Study of Automation and Accountability in Part-task Laboratory Simulation

FINAL REPORT

Cooperative Agreement NCC2-832

Western Aerospace Laboratories, Inc
Los Gatos, CA

Susan T. Heers
Principal Investigator

Submitted to:
NASA Ames Research Center
Moffett Field, CA
Everett A. Palmer, Ph.D., Technical Monitor

May 6, 1996

Paper presented at the *Eighth International Symposium on Aviation Psychology*,
Columbux, OH, April 1995. To be published in the proceedings

AUTOMATION AND ACCOUNTABILITY FOR PERFORMANCE

Kathleen L. Mosier
NASA Ames Research Center
San Jose State University Foundation

Linda J. Skitka
University of Illinois at Chicago

Susan T. Heers
NASA Ames Research Center
Western Aerospace Laboratories

Preliminary evidence suggests that rather than reducing error, automated procedural and decision aids may in some cases have the paradoxical effect of increasing it, or of promoting different errors. Recent research investigating the use of automated systems suggests that crews have a tendency to over-rely on automation to perform tasks and make decisions for them rather than using the aids as one component of thorough monitoring and decision-making processes, a phenomenon we are labeling *automation bias* (Mosier, Skitka, & Korte, 1994). Potential negative effects of automation bias can be broken down into two major types: (1) *automation commission errors,* i.e., errors made because crews take inappropriate action because they over-attend to automated information or direction, and (2) *automation omission errors,* i.e., errors made when crews do not take appropriate actions because they are not informed of an imminent problem or situation by automated aids.

A wide body of social psychological research has found that many cognitive biases and resultant errors can be ameliorated by imposing pre-decisional accountability, which sensitizes decision makers to the need to construct compelling justifications for their choices and how they make them. Accountability demands cause decision makers to employ more multidimensional, self- critical, and vigilant information seeking, and more complex data processing, and have been shown to reduce cognitive "freezing" or premature closure on judgmental problems (Kruglanski & Freund, 1983), and to lead decision makers to employ more consistent patterns of cue utilization (Hagafors & Brehmer, 1983). In the cockpit, accountability demands should lead to increased vigilance in decision making. Specifically, pilots should be less susceptible to automation bias, and more apt to check all relevant information before making decisions or taking actions.

To date, accountability effects have all been produced under controlled laboratory conditions and have not been demonstrated in a machine context. If similar shifts to more careful, data-based decision-making strategies can be produced in the cockpit, accountability may provide a means of decreasing automation bias and of promoting increased vigilance in decision making. Research into the effect of accountability on the use of automated decision aids needs to be conducted to determine the characteristics of any benefits in mitigating automation bias, as well as the potential trade-offs of imposing accountability (e.g., time factors).

This paper describes the first laboratory study in a program of research investigating the existence of automation bias in decision-making, and incorporating accountability demands as an ameliorative variable. Accountability for performance, as well as workload, and the reliability and ease of verification of an automated aid were manipulated in a complex, multi-task situation. Participants monitored and responded to gauge, text, and positional data, while performance a compensatory tracking task. The degree to which they followed the directives of an Automated Monitoring Aide (AMA), i.e., relied on the automation to dictate their responses without cross-checking with other, more reliable information, was a measure of automation bias. We expected this bias to be exacerbated by highly reliable automation, and attenuated by instructions concerning accountability for performance.

Methods

Experimental Tasks

Participants were 44 right-handed university students with normal or corrected to normal vision, who received course credit for their participation in the study. The primary experimental task was presented on a 13" color display monitor with four quadrants (see Figure 1), using the *Window /PANES: Workload/PerformANcE Simulation* (NASA-Ames Research Center, 1989). The top-left quadrant presented automated advisory messages for the button response tasks, and was referred to as the Automated Monitoring Aid (AMA). The top-right quadrant displayed the target and ownship symbols used in the compensatory tracking task. The lower-left quadrant displayed four dynamic, analog slide-rule gauges. Finally, the lower-right quadrant displayed a static top-down map which included a moving symbol depicting the target flight path position, used for the waypoint task.

Tracking Task. Participants performed a compensatory tracking task using the joystick with their right hand. They attempted to maintain alignment between the solid ownship symbol, which remained at the center of the quadrant, and the target circle, which moved about in the tracking window reflecting the disturbance imposed by a sum of sines algorithm. This task occurred continuously throughout the trial.

Gauge Advisories. The indicators in the four gauges changed their values in a pseudo-random manner throughout each of the trials. Subjects were instructed that the AMA automatically monitored the gauges, and if they needed to take actions to correct a pending "problem," the automation would "advise" them on what response to make. During a gauge event, a directive appeared in the AMA indicating a gauge problem and advising the subject on which of four response button to press (choice-response task). If three or four gauges were failing, the correct response would always be to press the button labeled "RESET." After a period of ten seconds, the directive left the screen, and the gauges moved back within their "safe" ranges until the next gauge event.

Waypoint Checks. A small square traversed the map throughout each trial depicting the subject's current target flight path position. Located on the map were several red or blue triangular symbols representing waypoint markers. Participants were instructed to "squawk" on a given frequency as they passed each waypoint marker, and were informed that the AMA would monitor this task and advise them when to "squawk." When the position marker passed the corner of a waypoint triangle, a directive appeared in the AMA alerting the subject to the waypoint check and identifying which frequency to use. Participants responded by pressing the appropriately labeled response button, then continued the trial. The directive left the screen after ten seconds.

Verification of Automated Directives. Participants were informed that the AMA was not 100% accurate and were provided with an auxiliary display to verify the automated directives. Participants were instructed that use of this system was optional, but that it would be necessary to achieve 100% accuracy. The verification program, which used the *MOUSELAB* decision laboratory software (Johnson, Payne, Schkade, & Bettman, 1991) was run on a second microcomputer and displayed on a second monitor. The program presented an eight-by-three matrix of cells (see Figure 2). Directives for the choice-response gauge tasks included a number from one to eight, corresponding to the appropriate row in the verification matrix. To verify a directive, participants located the row using the mouse with their left hand and clicked on the left-most cell. Once selected, the cell displayed information relevant to the gauge problem. Subjects were instructed to open as many cells as required by the directive (1-3, see Verification Level in the Design section) in order from left to right to completely verify the directive's accuracy. The final cell always contained the "correct" response (i.e., the response judged as accurate for feedback and performance measures).

Design and Procedures

The experiment was a mixed quasi-factorial design. The base experimental design manipulated Accountability and Reliability between participants and varied Tracking Difficulty within participants. Within each cell of this design, the discrete tasks represented two categories, each involving a different manipulation and a different aspect of automation bias. The choice-response gauge tasks provide a way to examine commission errors while assessing the effects of Ease of Verification. The waypoint and gauge reset tasks provided a means of examining omission errors while comparing predictable versus unpredictable monitoring tasks. Both of these manipulations were varied within participants. Response selections, response times, and tracking errors were analyzed as performance measures on the primary task. Additionally, data on when the participants used the auxiliary verification program were collected to evaluate how often and under what circumstances they double-checked the automation for the choice-response gauge tasks.

Participants were randomly assigned to one of two Accountability groups. Participants in the high accountability group were informed that their performance was being constantly monitored by the experimenter and that they would be asked to explain and justify their performance to the experimenters in an interview following the task. Participants in the non-accountable group were told that their performance could not be monitored by the experimenter, that their data were completely anonymous, and no mention of an interview was made.

Two levels of Automation Reliability were manipulated between subjects. In the high reliability condition, the AMA failed approximately 10% of the time, in the low reliability condition, approximately 40% of the time. Automation failures were represented by incorrect directives for the choice-response gauge tasks (the AMA indicated the wrong response button to press) or missed directives for the waypoint check and gauge reset tasks (the AMA failed to indicate that a response should be made). Automation failures were equally distributed across the discrete tasks. The Reliability manipulation was crossed with the Accountability factor, creating four between-subjects groups of eleven participants each.

For participants, half of the trials were performed under easy tracking conditions, and the other half were performed under difficult tracking conditions. Difficulty was manipulated by varying the bandwidth of the sum of sines algorithm imposed on the tracking axes. Three Verification levels defined the ease with which participants could cross-check the choice-response gauge tasks. The levels were defined by the number of informational items in the directive, either one, two, or three. The number of informational items corresponded to the number of boxes in the supplemental program matrix that the subject would have to open to completely verify the directive. Finally, two types of monitoring tasks were utilized to explore participants' reliance on automated advisories. The waypoint check task was a predictable monitoring task, since the subject was able to assess the distance between the position indicator and the waypoint marker throughout the trial and estimate when the task was about to occur. The gauge reset task was an unpredictable task, since the pseudo-random movement of the gauges prevented the participants from being able to anticipate when a reset condition (three or more gauges in red-zones) was likely to occur.

Participants were seated in a sound-dampened experimental booth with two monitors, the response box with the joystick and buttons, and a mouse. Participants performed a total of five training trials. First, participants performed two five-minute tracking-only trials under the two tracking difficulties. Participants then performed one four-minute trial under easy tracking in which discrete task events representing each of the task conditions occurred every 35-50 seconds. Following this, participants in the reliable condition were told that the automation had been tested and found to be highly reliable, and those in the unreliable condition were told that the reliability of the system was uncertain. All participants were informed that the verification display was 100% accurate, and that use of this program was optional and at their discretion. Participants then completed the final two five-minute practice trials, one for each tracking difficulty. In both of these trials, tasks occurred every 45-55 seconds, and the experimenter encouraged the subjects to use the verification program.

Following a short break, the experimenter read the instructions corresponding to the subject's accountability condition (see the Design section). Participants then performed a total of six experimental trials, with a short break midway through the experiment. Within each tracking difficulty condition, the choice-response gauge tasks were replicated eight times under each of the three verification levels, for a total of 24 tasks. The waypoint and gauge reset tasks were replicated four times each. For each tracking condition, these 32 events were distributed across three trial scenarios, with two of the trials containing 11 task events and one trial containing 10 task events. In all trials, events occurred every 50 ±10 seconds and were randomly ordered. Measures for each task condition were collapsed across replications within each tracking condition for purposes of analysis.

At the end of each trial, participants were given visual feedback on their mean response time, percentage correct, and root mean squared tracking error. Following the last trial, subjects completed a short questionnaire which proved their attitudes toward the task and toward computers in general, and also contained items designed to verify the success of the accountability manipulation. The experimenter debriefed the subject as to the purposes of the study, and any questions were answered. None of the participants received the justification interview described in the accountability condition.

Results

The results dealing with the choice-response task, verification behavior, and questionnaire responses will be summarized in this paper. Other performance results have been discussed in a previous paper (Heers, Marchioro, Mosier, & Skitka, 1994). Validating the experimental manipulations, accountable participants were significantly more likely to report that the experimenter was monitoring their performance than those who were not accountable [$p<.01$], and to report that their performance was being evaluated [$p<.01$].

For the choice-response tasks, response selection, verification, and response time data for the choice-response gauge tasks were each analyzed using a 2 x 2 x 2 x 3 (Accountability x Reliability x Tracking Difficulty x Verification Level) mixed factorial Analysis of Variance (ANOVA) design.
Analysis of the number of times participants selected the correct response showed that participants in the reliable condition were more likely to make a correct response (91.57%) than participants in the unreliable condition (83.71%) [M.E. Reliability: $F(1, 40) = 7.01$, $p = .012$]. This difference is in same direction as the expected accuracies if participants followed the AMA directives without verifying, since participants in the unreliable condition would be given more incorrect responses than those in the reliable condition. The observed accuracies are higher for the two groups than the accuracy of the automated directives each group received, however, implying that subjects in both groups verified the automation to some degree. For example, if participants in the unreliable group did not verify any of the directives, they would only be expected to achieve a 62.5% accuracy, the same as the reliability of their automation.

Verification Behavior. The use of the auxiliary verification program served as the measure of the cross-checking behavior of participants. The number of times subjects completely verified a directive was summed across the eight replications within each cell of the design and analyzed. These results showed while there was an overall decrease in verifications under difficult tracking conditions [M.E. Tracking: F (1, 40) = 6.89, p = .012], the two accountability groups responded differently to the tracking manipulation [Accountability x Tracking: F (1, 40) = 4.25, p = .046]. Accountable participants were more likely to verify under easy (4.50) than difficult (3.74) tracking, while non-accountable participants showed virtually the same verification behavior under easy (5.20) and difficult (5.11) conditions.

Participants verified to a lesser degree as the number of boxes required to verify the directive increased [M.E. Verification: F (2, 80) = 8.94, p < .001]. This trend was modified by the interaction of verification level with accountability [F (2, 80) = 3.75, p = .028]. While the accountable group did show a significant drop in the number of verifications between the first level and the second and third levels (see Figure 3), the non-accountable group had statistically equivalent numbers of verifications across the three levels. This suggests that the accountable subjects were more sensitive to the increasing demands of the higher verification levels and chose to verify the longer directives less often to avoid other costs to performance.

These results were complemented by the results of the analysis on the response times, indicating that the use the verification program did produce the expected trade-off in the speed with which subjects were able to respond to the discrete tasks. Subjects took longer to respond to the task as the verification level increased [F (2, 80) = 33.53, p < .001], reflecting the number of boxes subjects were required to open. In addition, the data patterns also complement the interaction with the Accountability manipulation observed in the verification data. As seen in Figure 4, the response times for the accountable group did not increase as much across verification levels as for the non-accountable group, with level three not reliably different from level two. This smaller increase in the time for level three would be expected given the lower number of verifications performed by the accountable group in the level three condition.

Finally, the verification data showed an interaction between reliability, tracking difficulty, and verification level [F (2, 80) = 4.03, p = .021]. As can be seen in Figure 5, the reliable subjects showed a consistent decrease in the number of verifications completed from the first verification level to the second and third levels in both tracking difficulties. These subjects also showed an overall decrease in the number of verifications completed as the tracking difficulty increased. The pattern for the unreliable subjects, on the other hand, indicate that under easy tracking conditions, the verification levels were statistically the same. Subjects in this condition only began to decrease this behavior under the most demanding conditions, when both tracking and verification difficulty were high. This suggests that subjects with unreliable directives were more reluctant to give up verifying the AMA accuracy as other demands increased.

The data on the response accuracy and verification behavior support the presence of an automation bias in this study. Subjects tended to rely on the directives to provide them with information on which responses to make, and chose to verify these directives an average of only 58% of the time. The reliability of the automation had a significant impact on this bias. While subjects in the unreliable condition still performed worse on all of the discrete tasks, their performance was much higher than that of the automation itself, while subjects in the reliable condition were only marginally better than if they had accurately followed the automation without verifying. Automation reliability appears to have affected the strategies adopted by the two groups regarding verification. As the time required to verify a directive increased, subjects in the reliable condition, who encountered very few errors in the directives, were more likely to shed the verification task. The subjects in the unreliable condition, on the other hand, were reluctant to decrease their verification behavior as demands increased. Only when the increase in verification time was coupled with more difficult tracking did the subjects in the unreliable condition begin to shed the verification behavior.

The nature of the accountability effects in this experiment were unanticipated. The accountable subjects appeared to have adopted a strategy similar to that observed by the subjects in the reliable condition. As the demands of the verification task increased, subjects in the accountable condition were more likely to shed the verification task. This resulted in better response times for the accountable group relative to the non-accountable group. Rather than increasing verification behavior, then, accountability had the effect of prompting subjects to maximize speed at the expense of accuracy.

To further examine the factors that influenced verification behavior, correlations were calculated between the verification variables and responses to items on the questionnaires. (The following relationships are statistically significant at at least the .05 level.) Subjects who reported that they were nervous or that they were aware of being monitored were significantly more likely to stop verifying as it became more difficult (e.g., when tracking was difficult, or when verification required opening two or three boxes). Attitudes towards computers as well as the age of the participants also influenced verification behavior, as participants who were older, or who thought computers were "perfect" or "objective" were less likely to verify. Those participants with more video-game experience were more likely to complete the most difficult verifications, possibly reflecting more ease with the task in general. Interestingly, participants who thought that the automation should receive credit for their monitoring and response performance were also those who were more likely to verify the automated directives.

The presence of automation bias in this study and its interactions with reliability and ease of verification indicate a need to further investigate the characteristics of this phenomenon and its interaction with other relevant variables in both laboratory and naturalistic settings. Additionally, the accountability manipulations in this study showed an impact on the strategies adopted by subjects in dealing with increased task demands. Since all performance aspects of the task were given equal priority, it is assumed that the subjects made their own determination as to how best to optimize performance. As discussed, this often entailed sacrificing response accuracy for high performance on the other metrics. A second study, now in progress, will more precisely define the aspect of performance for which participants will be held accountable. This will allow us to more directly link accountability to specific task performance. Additionally, more research needs to be conducted to determine task and environmental variables which interact with accountability demands in a real world environment.

Figures

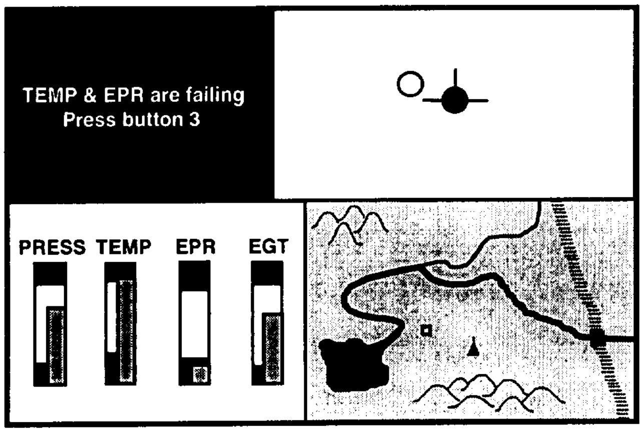

Figure 1: Primary task display.

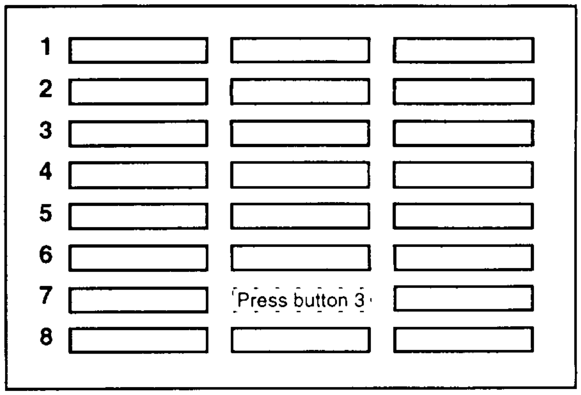

Figure 2. Auxiliary Verification Program Display

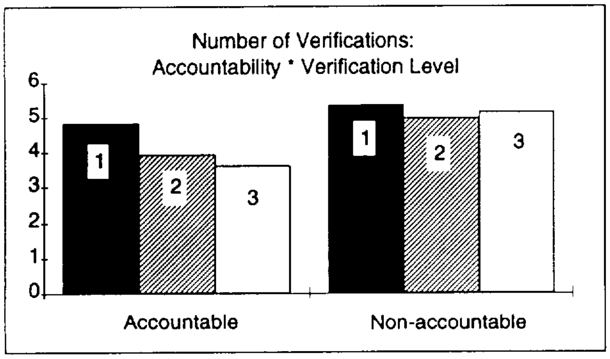

Figure 3. Number of verifications for each accountability group across verification levels.

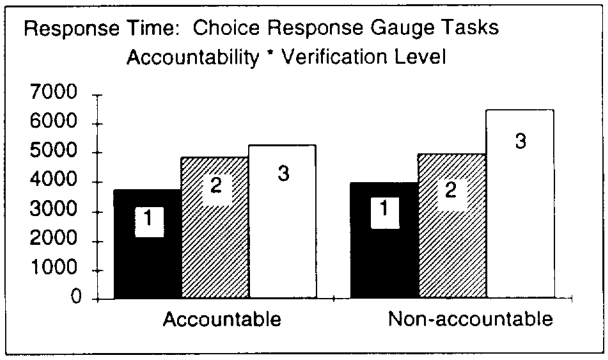

Figure 4. Response times for choice-response gauge tasks for each accountability group across verification levels.

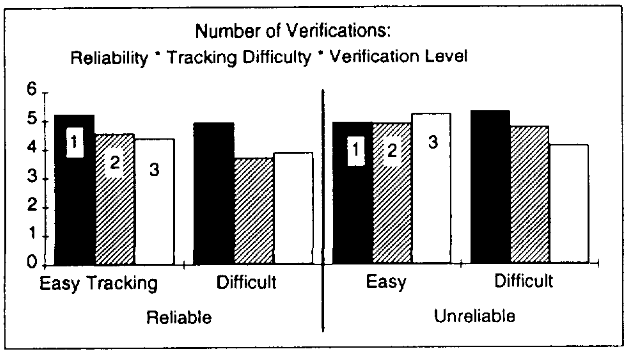

Figure 5. Number of verifications for each reliability group by tracking difficulty.

References

Hagafors. R., & Brehmer, B. (1983). Does having to justify one's decisions change the nature of the decision process? *Organizational Behavior and Human Performance, 31*, 223-232.

Heers, S. T., Marchioro, C. A., Mosier, K. L., & Skitka, L. J. (1994). *Automation and accountability in a low fidelity flight task.* Poster presented at the 38th Annual Meeting of the Human Factors and Ergonomics Society, Nashville, TN.

Johnson, E. J., Payne, J. W., Schkade, D. A., & Bettman, J. R. (1991). *Monitoring information processing and decisions: The MouseLab system.* Philadelphia: University of Pennsylvania, The Wharton School.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology, 14*, 448-468.

Mosier, K. L., Skitka, L. J., & Korte, K. J. (1994). Cognitive and social psychological issues in flight crew/automation interaction. *Proceedings of the Automation Technology and Human Performance Conference*, Sage.

NASA Ames Research Center (1989). Window/PANES: Workload PerformANcE Simulation. Moffett Field, CA: NASA Ames Research Center, Rotorcraft Human Factors Research Branch.